Epigenetics and DNase-Seq

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2017

Anthony Gitter

gitter@biostat.wisc.edu

Goals for lecture

Key concepts

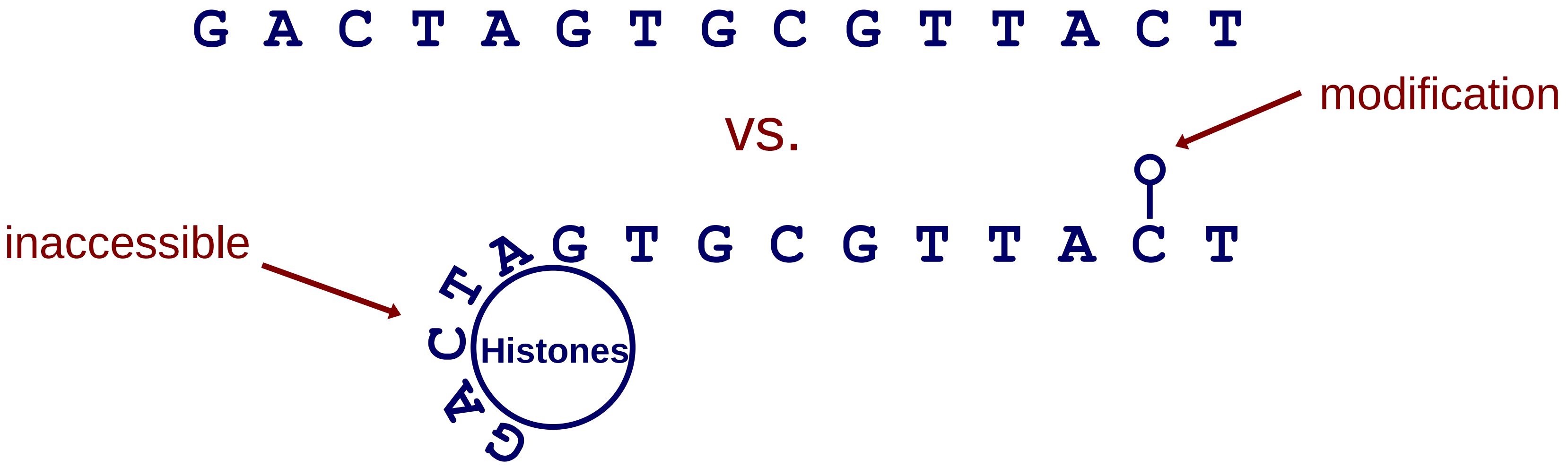
- Importance of epigenetic data for understanding transcriptional regulation
- Predicting transcription factor binding sites
- Gaussian Process models

Introduction to epigenetics

Defining epigenetics

- Formally: attributes that are “in addition to” genetic sequence or sequence modifications

- Informally: experiments that reveal the context of DNA sequence
  - DNA has multiple states and modifications



G A C T A G T G C G T T A C T

vs.

G A C T A G T G C G T T A C T

Importance of epigenetics

Better understand

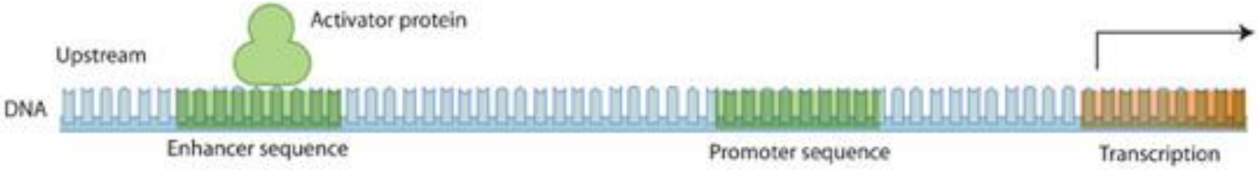
- DNA binding and transcriptional regulation
- Differences between cell and tissue types
- Development and other important processes
- Non-coding genetic variants (next lecture)

PWMs are not enough

- Genome-wide motif scanning is imprecise
- Transcription factors (TFs) bind < 5% of their motif matches
- Same motif matches in all cells and conditions

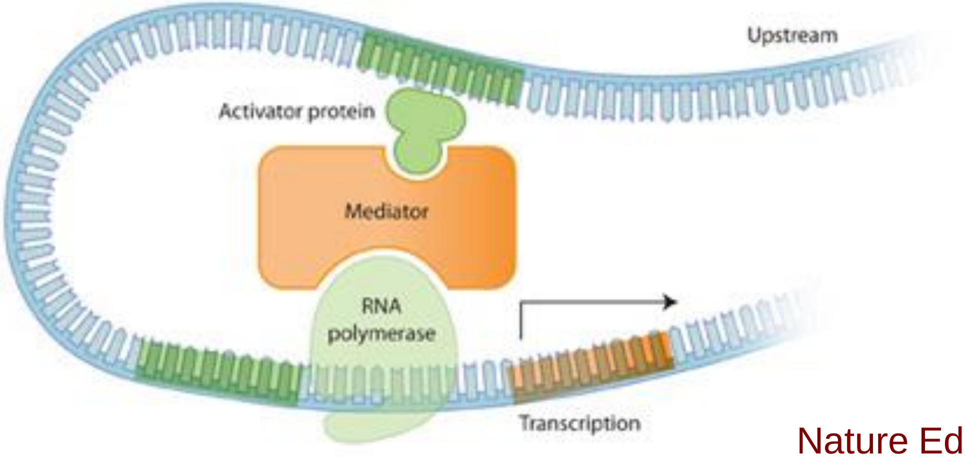
PWMs are not enough

- DNA looping can bring distant binding sites close to transcription start sites
- Which genes does an enhancer regulate?



Promoter: DNA binding site for TFs, close to gene transcription start site

Enhancer: DNA binding site for TFs, can be far from affected gene

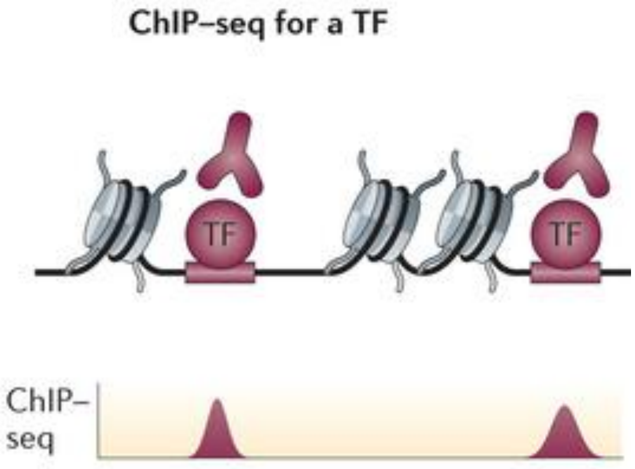


Nature Education 2010

Mapping regulatory elements genome-wide

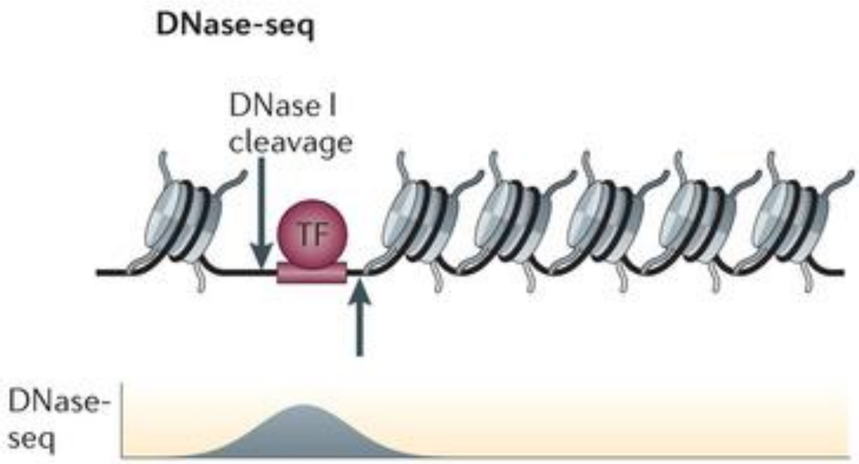
- Can do much better than motif scanning with additional data

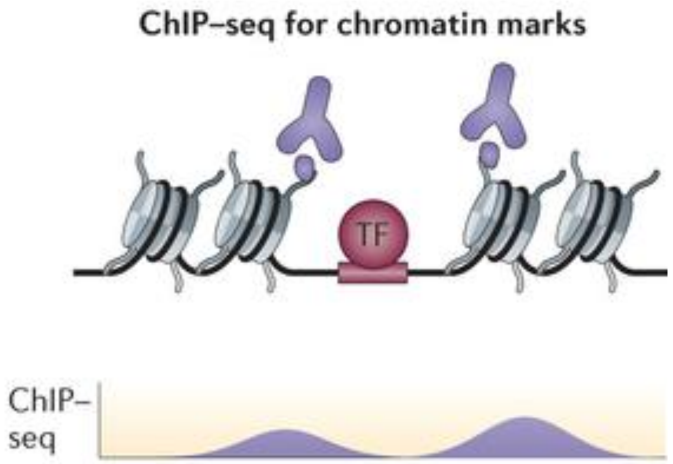
- ChIP-seq measures binding sites for one TF at a time



Shlyueva *Nature Reviews Genetics* 2014

- Epigenetic data suggests where *some* TF binds

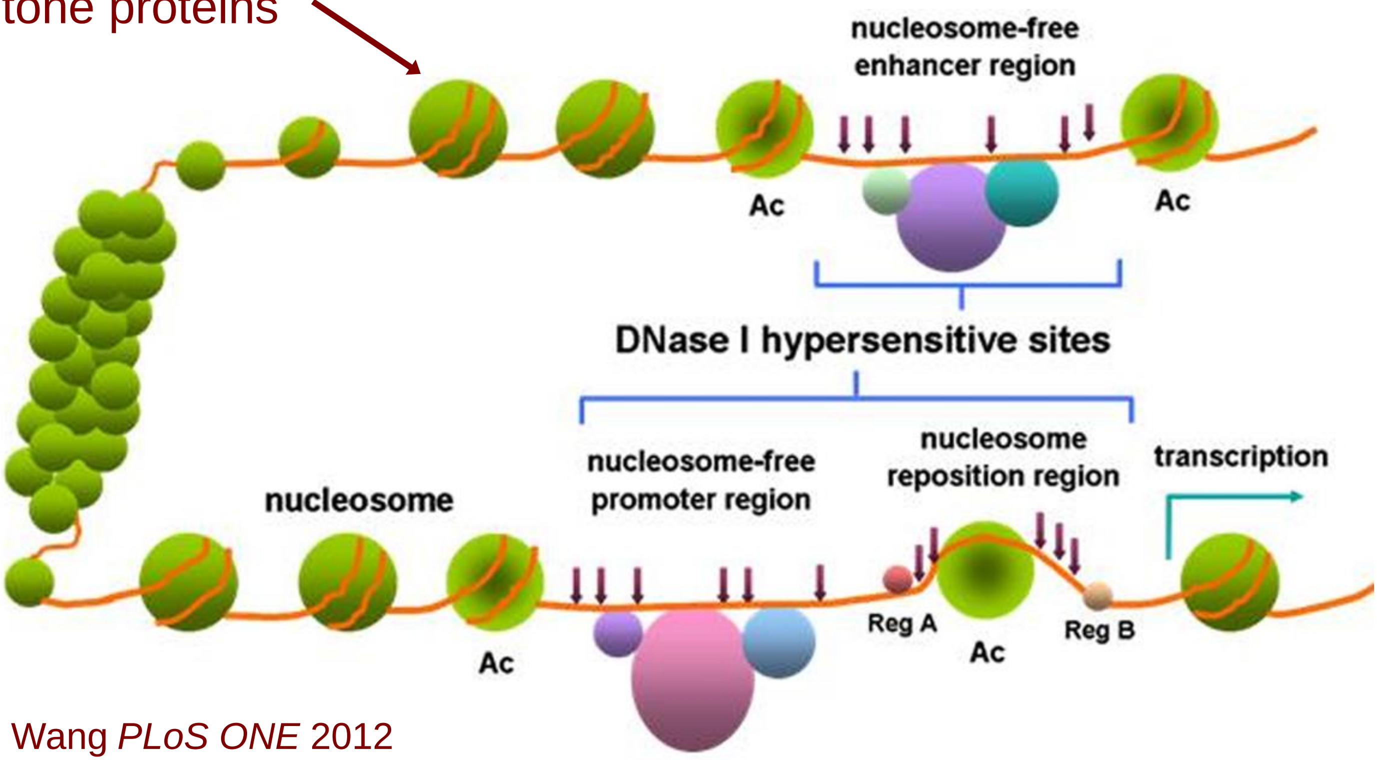




DNase I hypersensitivity

- Regulatory proteins bind accessible DNA
- DNase I enzyme cuts open chromatin regions that are not protected by nucleosomes

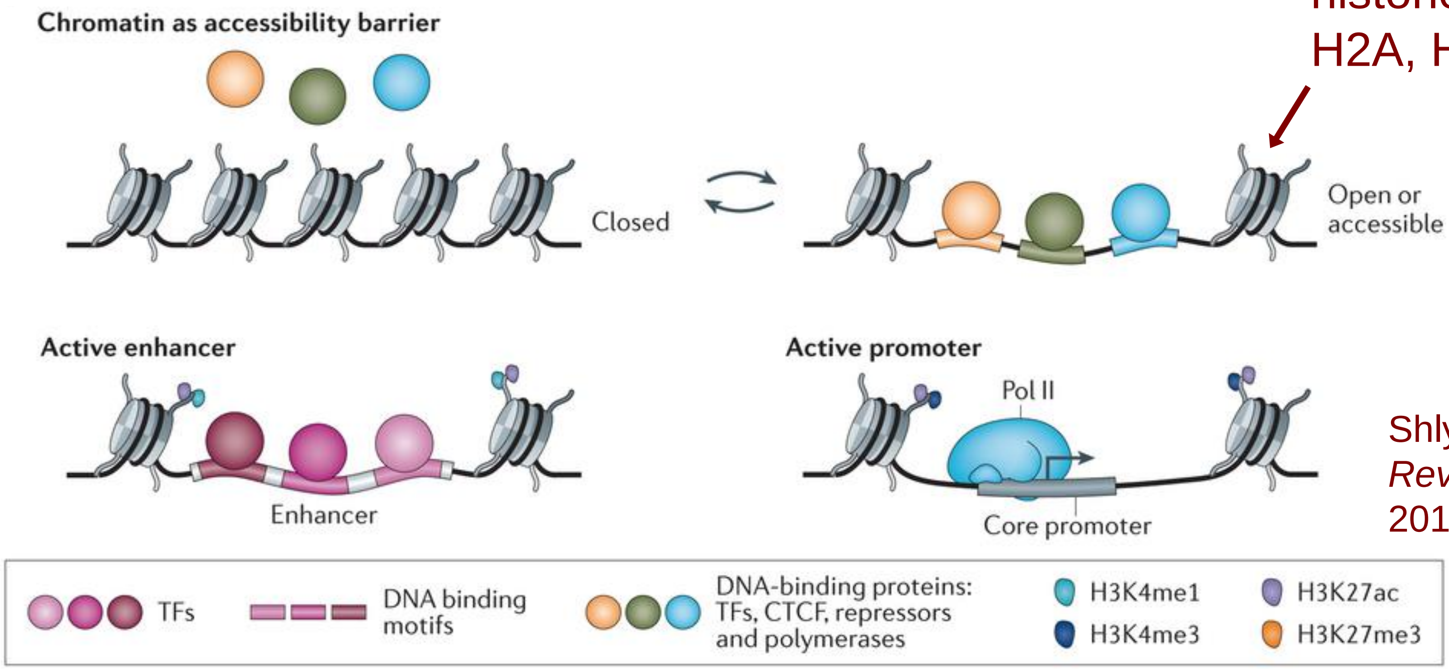
Nucleosome: DNA wrapped around histone proteins



Wang *PLoS ONE* 2012

Histone modifications

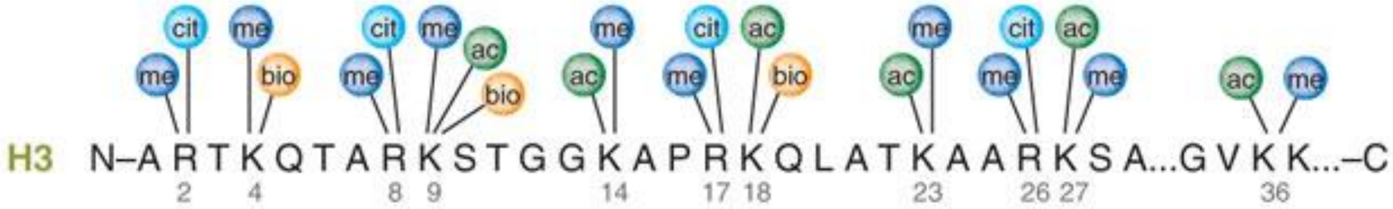
- Mark particular regulatory configurations



Two copies of histone proteins H2A, H2B, H3, H4

Shlyueva *Nature Reviews Genetics* 2014

- H3 (protein) K27 (amino acid) ac (modification)

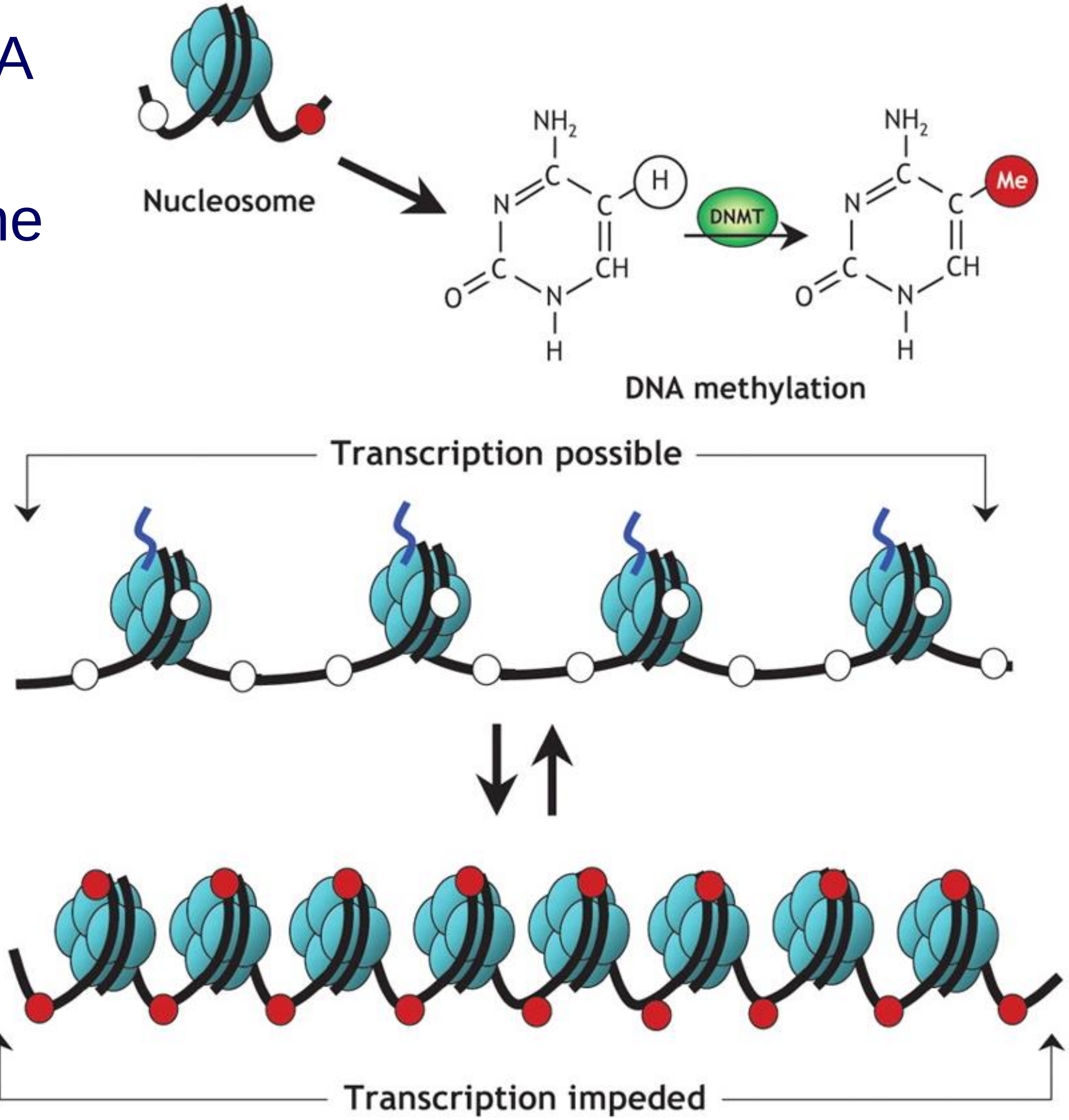


H3
N–A R T K Q T A R K S T G G K A P R K Q L A T K A A R K S A...G V K K...–C
2 4 8 9 14 17 18 23 26 27 36
cit
me
bio
ac

Latham *Nature Structural & Molecular Biology* 2007; Katie Ris-Vicari

DNA methylation

- Reversible DNA modification
- Represses gene expression



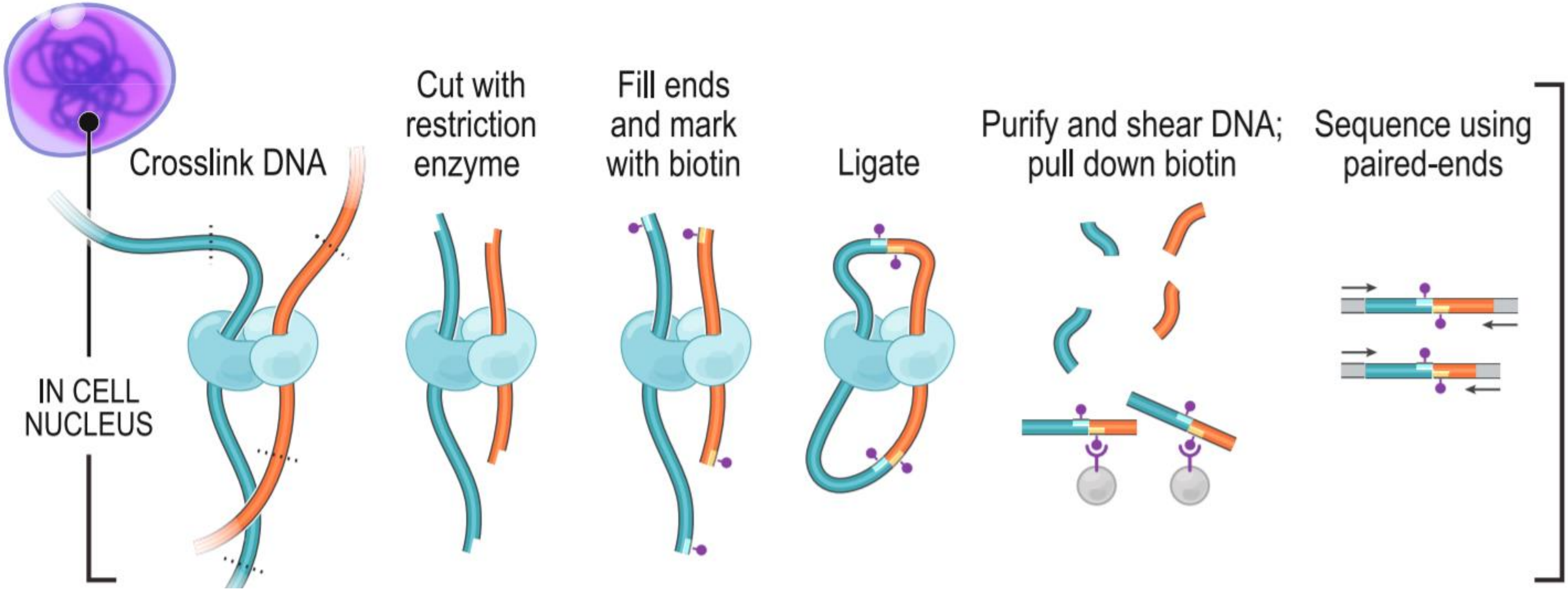
Gene “switched on”
- Active (open) chromatin
- Unmethylated cytosines (white circles)
- Acetylated histones

Gene “switched off”
- Silent (condensed) chromatin
- Methylated cytosines (red circles)
- Deacetylated histones

OpenStax CNX

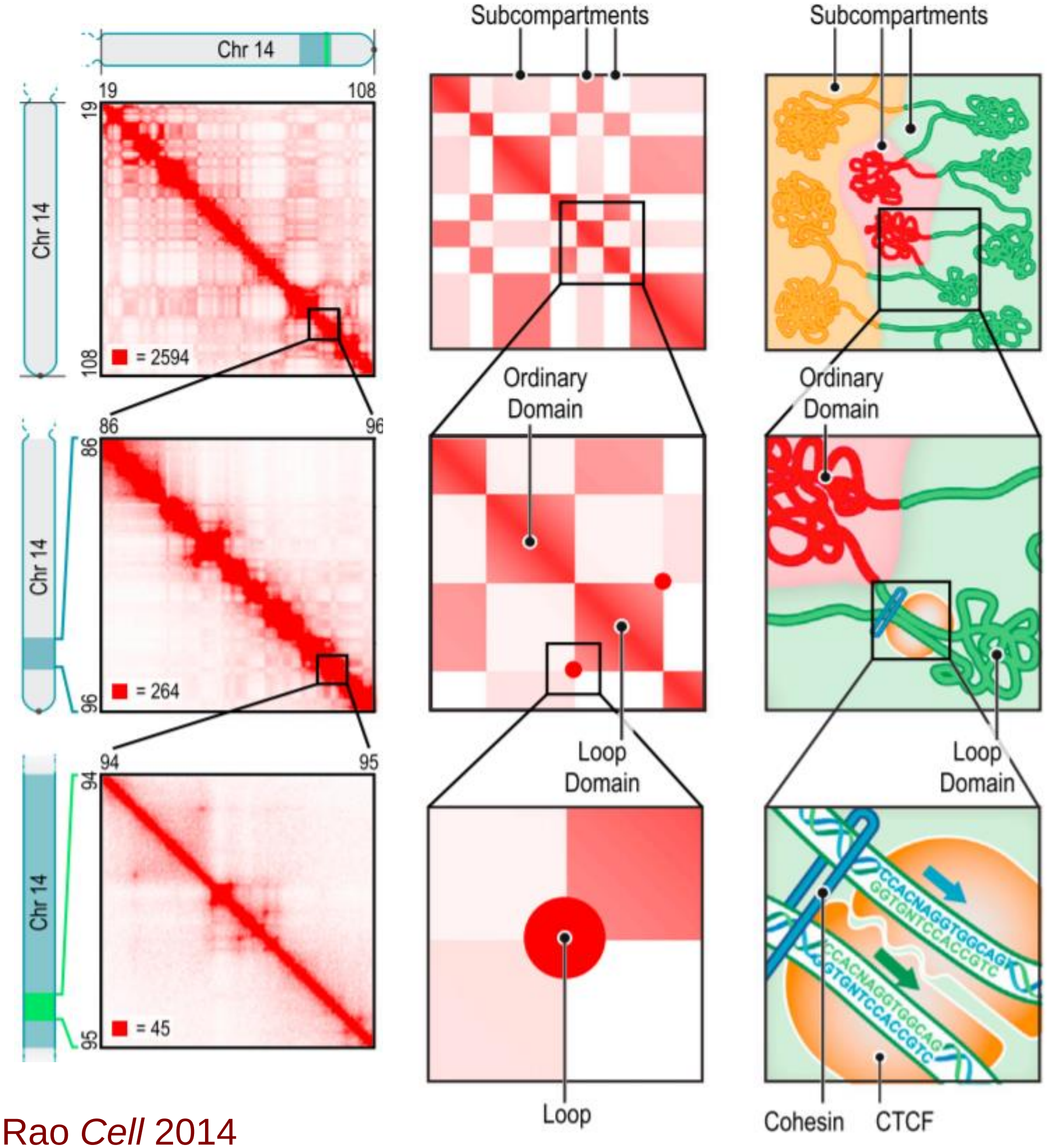
3d organization of chromatin

- Algorithms to predict long range enhancer-promoter interactions
- Or measure with chromosome conformation capture (3C, Hi-C, etc.)


Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
IN CELL NUCLEUS

Rao *Cell* 2014

3d organization of chromatin



500 kb

50 kb

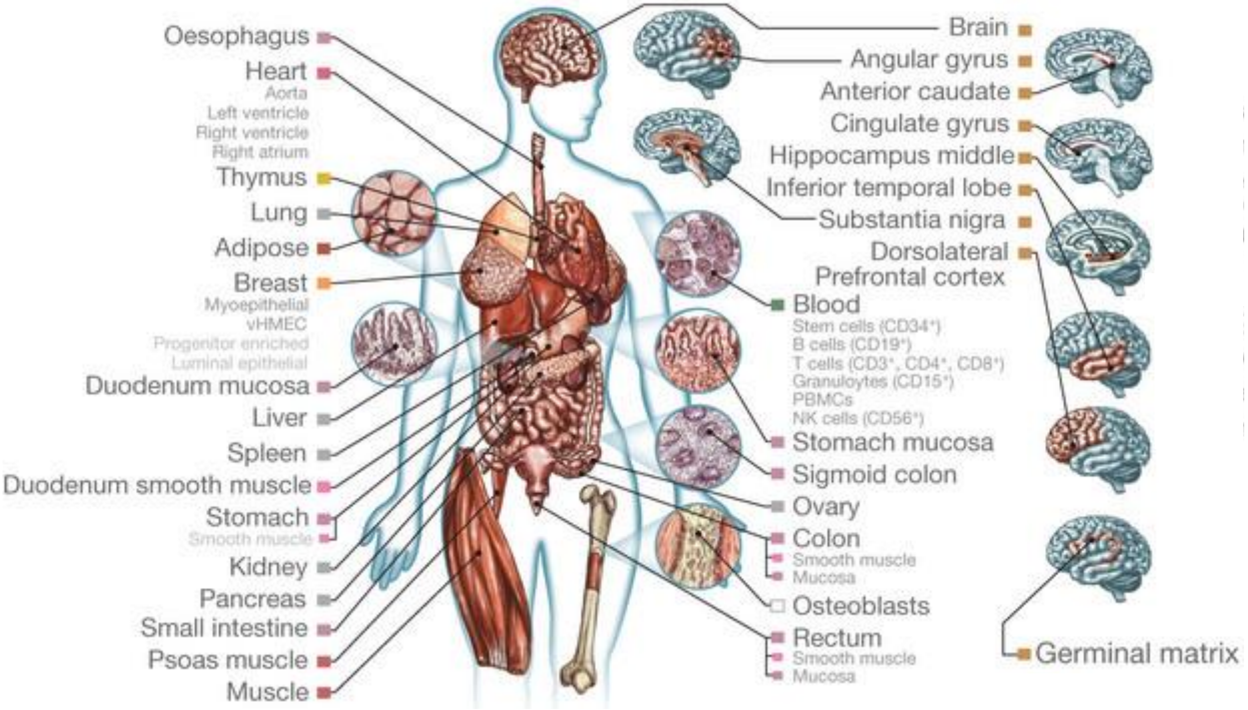
5 kb

Rao *Cell* 2014

- Hi-C produces 2d chromatin contact maps
- Learn domains, enhancer-promoter interactions

Large-scale epigenetic maps

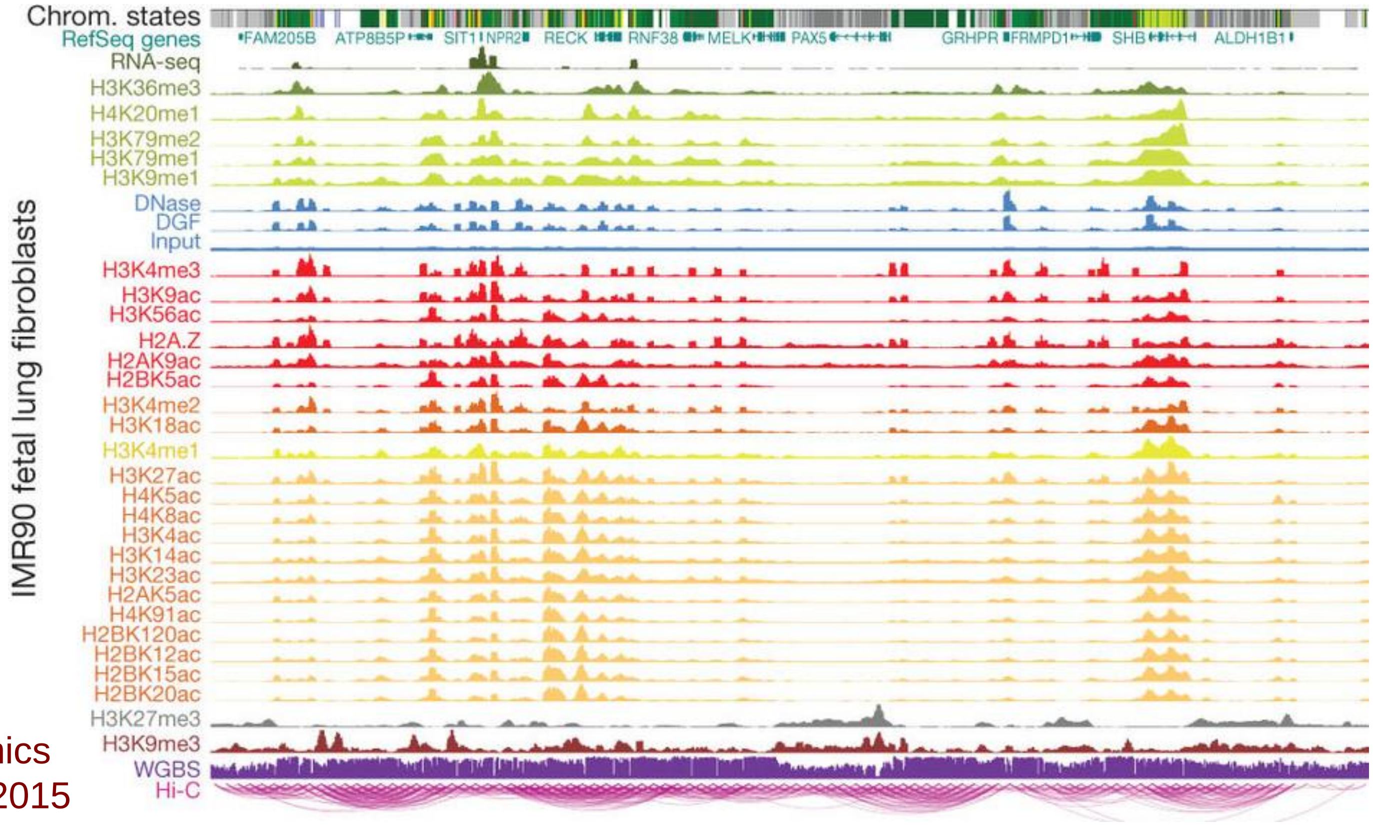
- Epigenomes are condition-specific
- Roadmap Epigenomics Consortium and ENCODE surveyed over 100 types of cells and tissues



Roadmap Epigenomics Consortium *Nature* 2015

Genome annotation

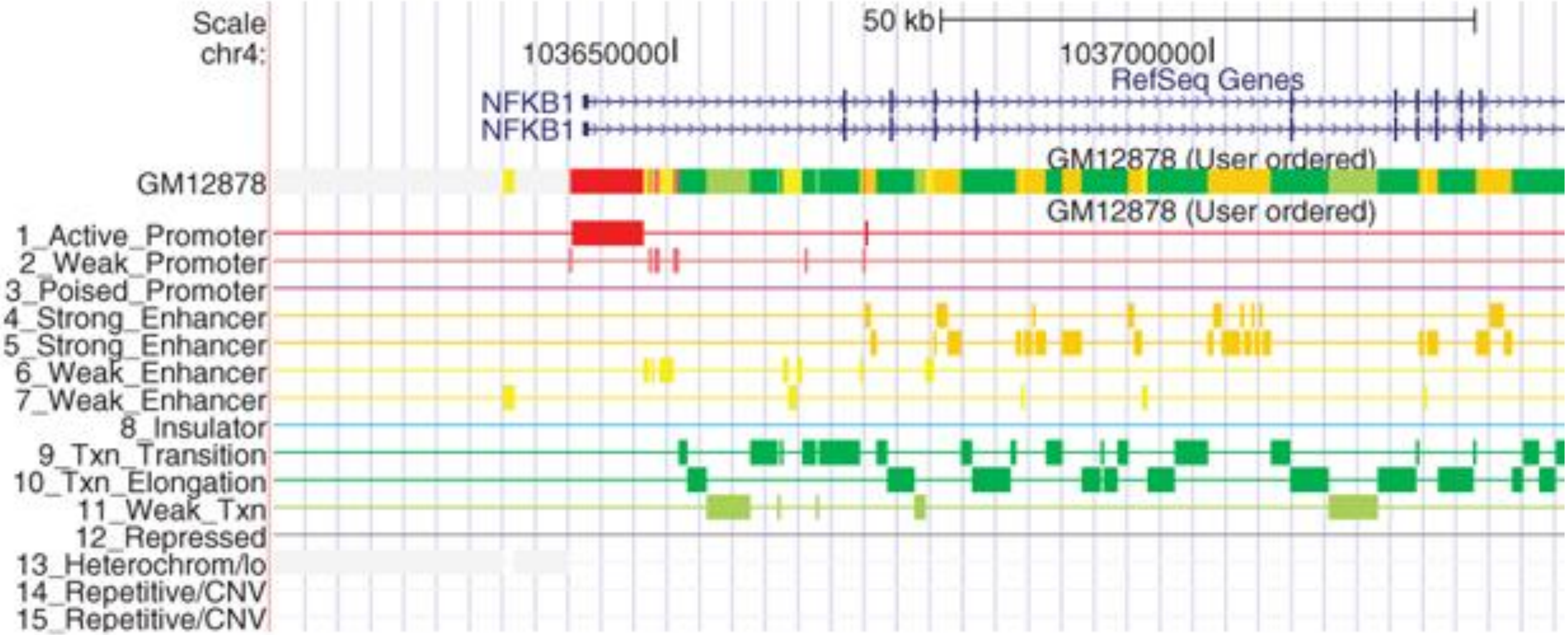
- Combinations of epigenetic signals can predict functional state
  - ChromHMM: Hidden Markov model
  - Segway: Dynamic Bayesian network



Roadmap Epigenomics Consortium *Nature* 2015

Genome annotation

- States are more interpretable than raw data

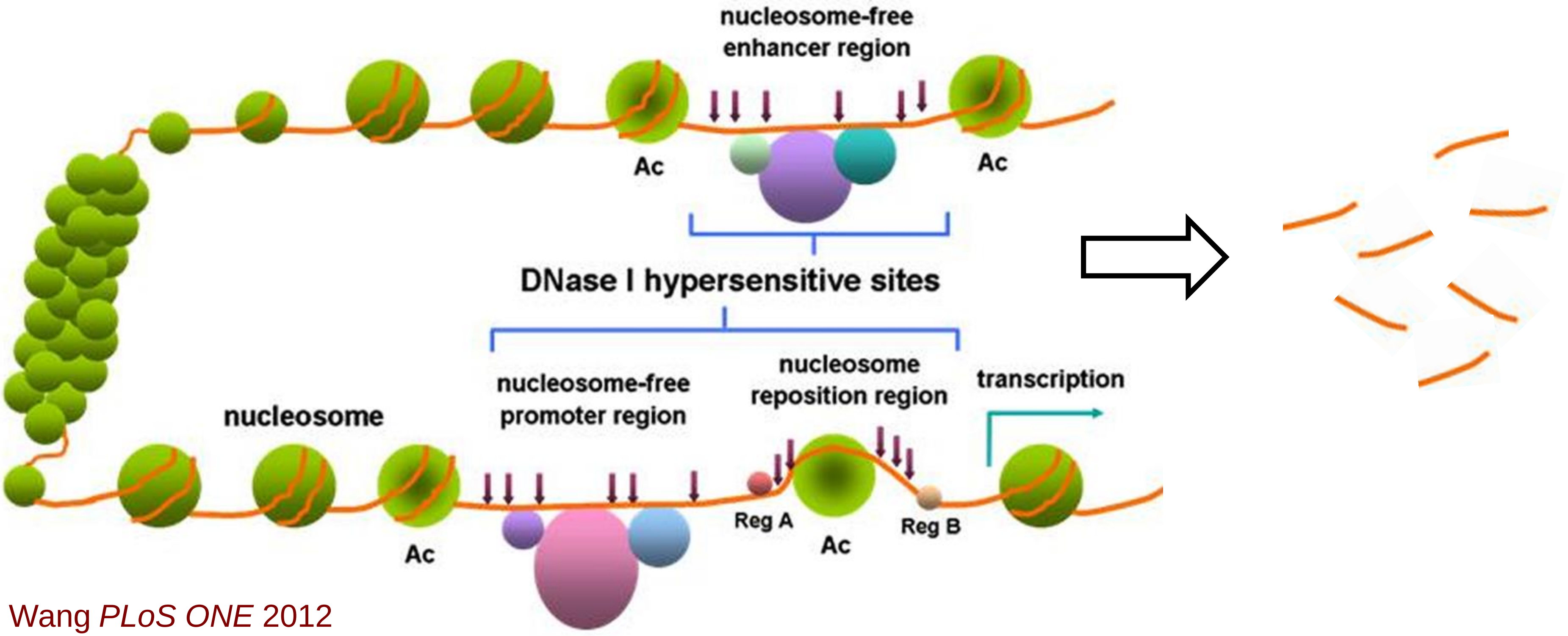


Ernst and Kellis *Nature Methods* 2012

Predicting TF binding with DNase-Seq

DNase I hypersensitive sites

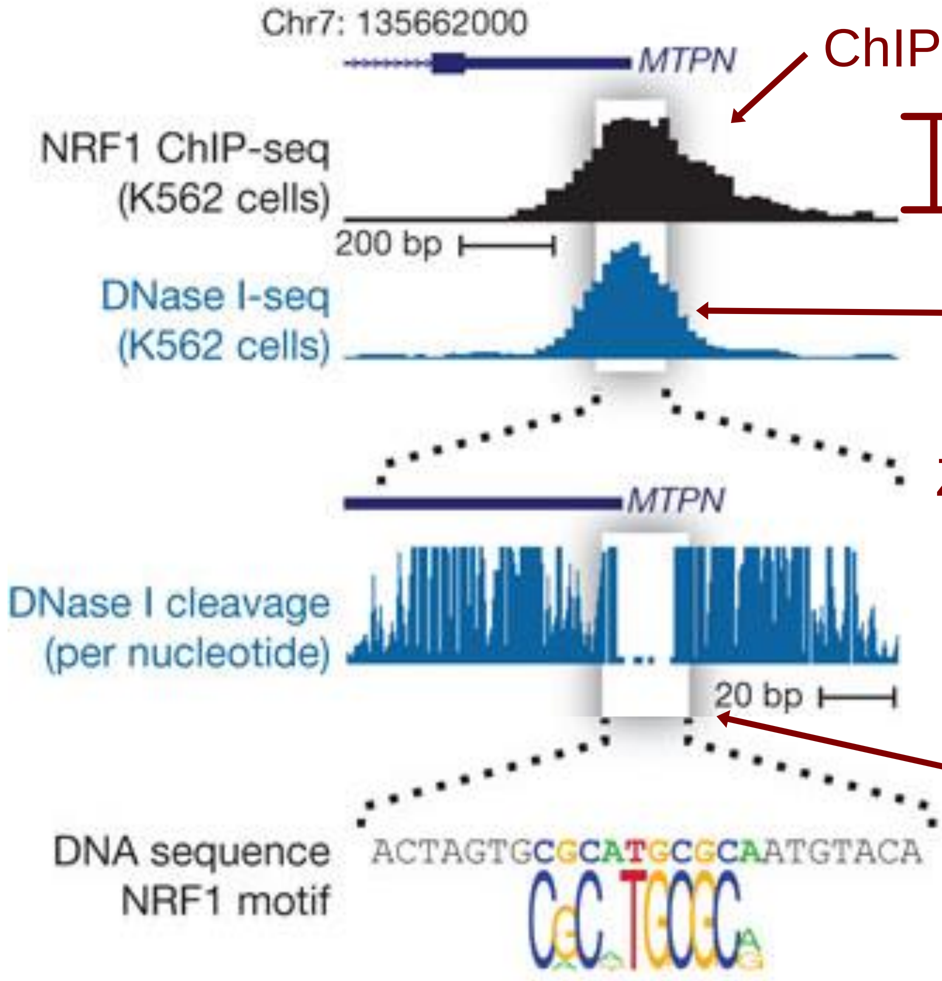
- Arrows indicate DNase I cleavage sites
- Obtain short reads that we map to the genome



Wang *PLoS ONE* 2012

DNase I footprints

- Distribution of mapped reads is informative of open chromatin and specific TF binding sites



ChIP-Seq peak

Read depth at each position

Nucleosome free “open” chromatin

Zoom in

TF binding prevents DNase cleavage leaving Dnase I “footprint”, only consider 5′ end

Neph *Nature* 2012

DNase I footprints to TF binding predictions

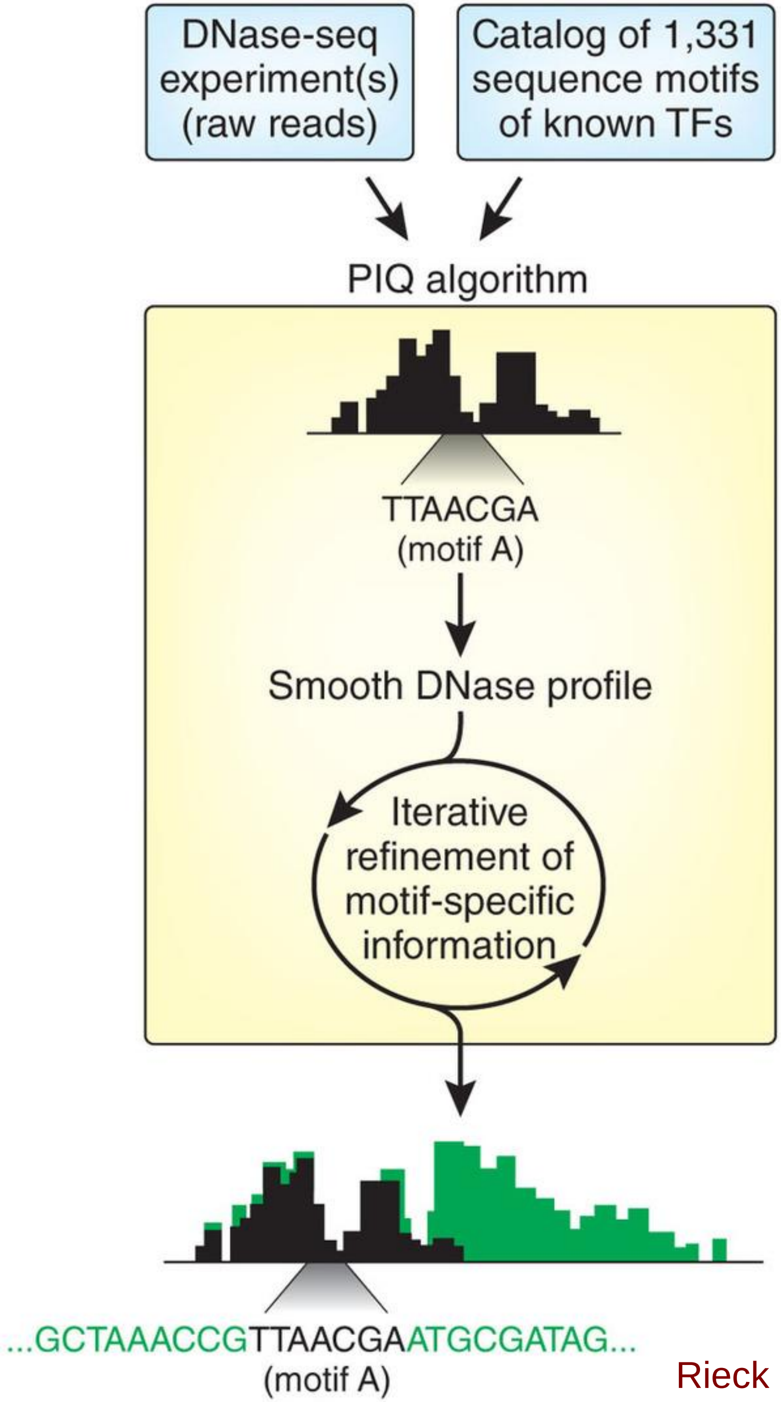
- DNase footprints suggest that ***some*** TF binds that location

- We want to know ***which*** TF binds that location

- Two ideas:
  - – Search for DNase footprint patterns, then match TF motifs
  - – Search for motif matches in genome, then model proximal DNase-Seq reads

We'll consider this approach

Protein Interaction Quantification (PIQ)


DNase-seq experiment(s) (raw reads)
Catalog of 1,331 sequence motifs of known TFs
PIQ algorithm
TTAACGA
(motif A)
Smooth DNase profile
Iterative refinement of motif-specific information
...GCTAAACCGTTAACGAATGCGATAG...
(motif A)

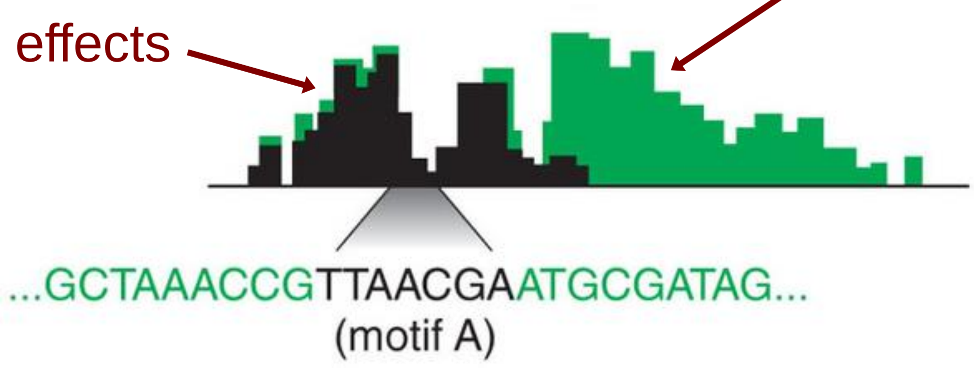
- Sherwood et al. *Nature Biotechnology* 2014
- **Given**: TF motifs and DNase-Seq reads
- **Do**: Predict binding sites of each TF

Rieck and Wright *Nature Biotechnology* 2014

PIQ main idea

- With no TF binding, DNase-Seq reads come from some background distribution

- TF binding changes read density in a *TF-specific* way

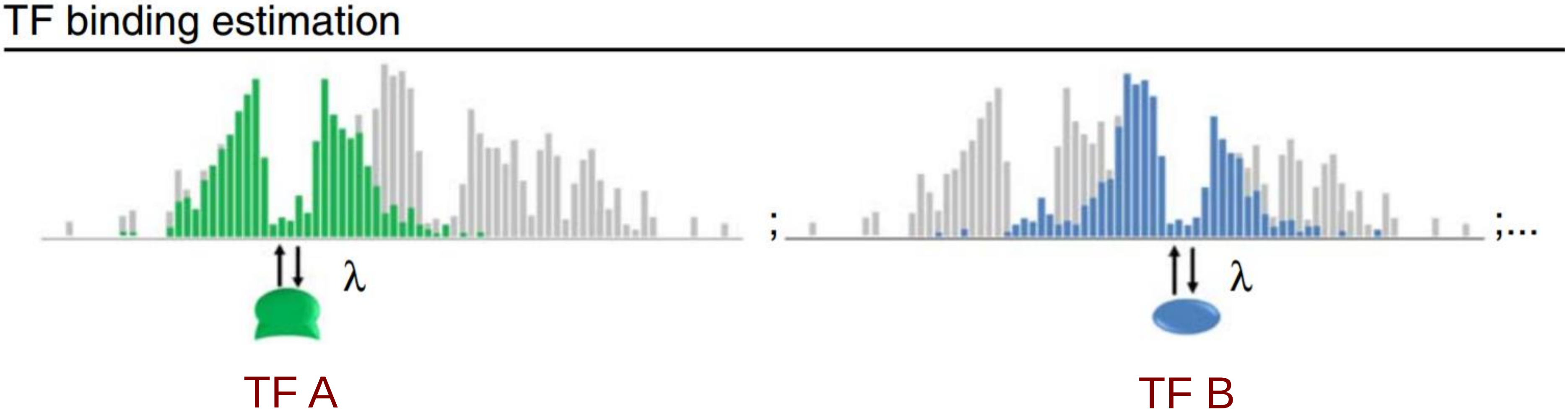


Background

TF effects

PIQ main idea

- Shape of DNase peak and footprint depend on the TF



Sherwood *Nature Biotechnology* 2014

PIQ features

- We’ll discuss
  - Modeling the DNase-Seq background distribution
  - How TF binding impacts that distribution
  - Priors on TF binding

- We’ll skip
  - Modeling multiple replicates or conditions, cross-experiment and cross-strand effects
  - Expectation propagation
  - TF hierarchy: pioneers, settlers, migrants

Algorithm preview

- Identify candidate binding sites with PWMs
- Build a probabilistic model of the DNase-Seq reads
- Estimate TF binding effects
- Estimate which candidate binding sites are bound
- Predict pioneer, settler, and migrant TFs

DNase-Seq background

- Each replicate is noisy, don’t want to over-interpret this noise
  - Only counting density of 5′ ends of reads

- Manage two competing objectives
  - Smooth some of the noise
  - Don’t destroy base pair resolution signal

Gaussian processes

- Can model and smooth sequential data
- Bayesian approach
- Jupyter notebook demonstration

TF DNase profile

- Adjust the log-read rate by a TF-specific effect at binding sites

DNase profile for factor *l*

Whether site *m* is bound

$$\widehat{\mu_l} = \mu_i + \begin{cases} \beta_{i-j,l} & |y_m - j| \le W \mathit{\ and\ } I_m = 1 \\ 0 & \mathit{otherwise} \end{cases}$$

DNase log-read rate adjusted for binding of factor *l*

Location of binding site *m*

Window size

DNase log-read rate at position *i* from Gaussian process

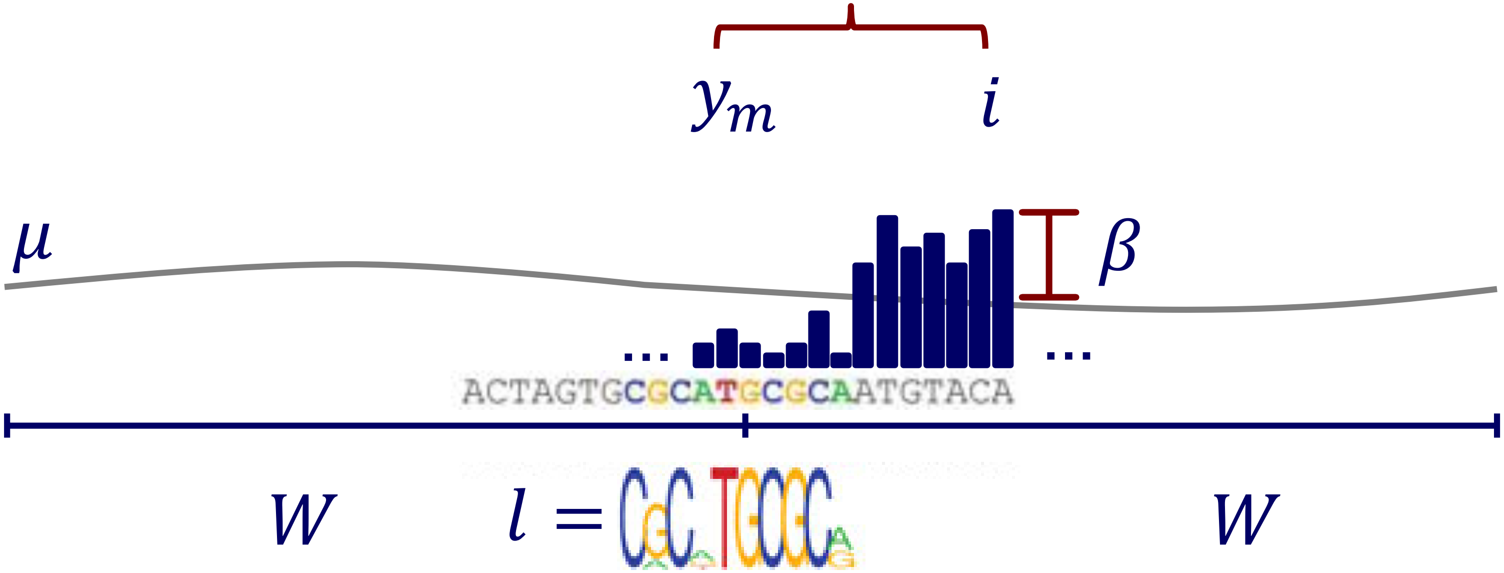
TF DNase profile

- DNase profiles represented as a vector for each TF

DNase profile for factor *l*

$$\widehat{\mu_l} = \mu_i + \begin{cases} \beta_{i-j,l} & |y_m - j| \le W \; and \; I_m = 1 \\ 0 & otherwise \end{cases}$$

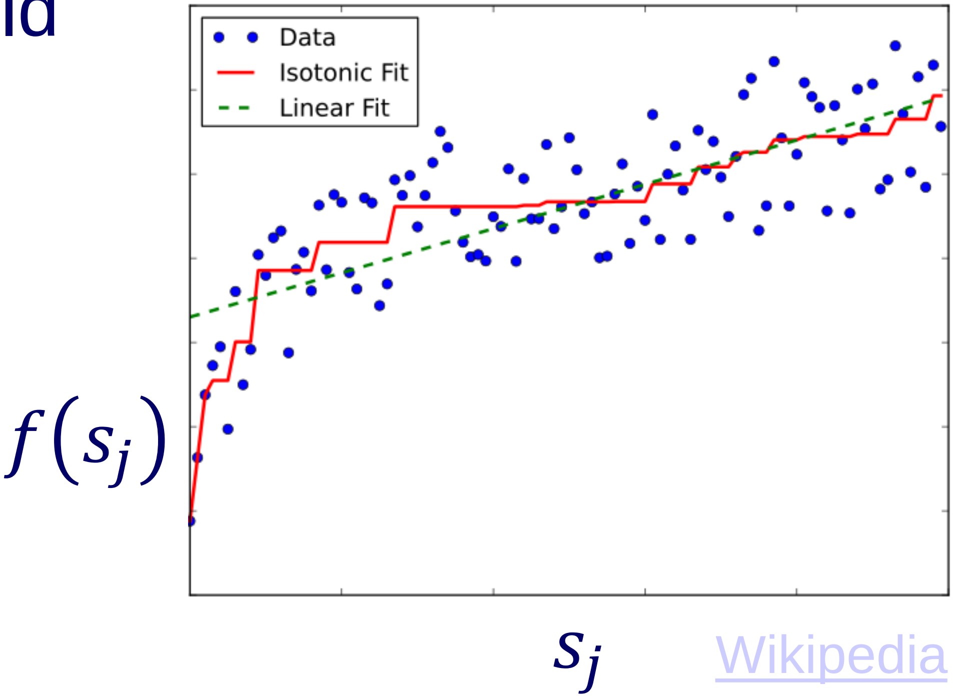
Can't be too far apart



Priors on TF binding

- TF binding event I_j should be more likely when
  - motif score s_j is high
  - DNase counts c_j are high
- Isotonic (monotonic) regression

Example only, not realistic data



Wikipedia

$$\log(P(I_j = 1)) = f(s_j) + g(c_j)$$

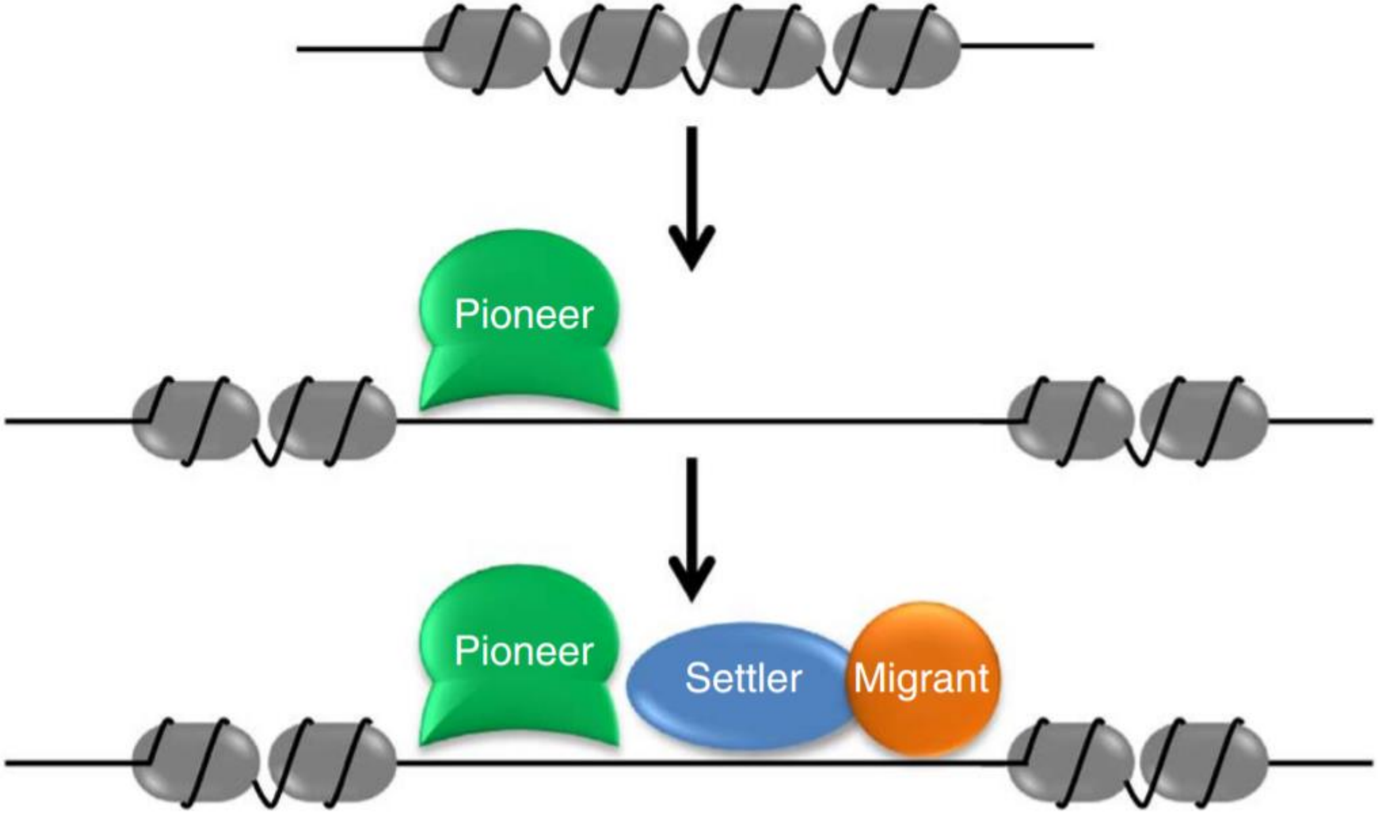
Full algorithm

- **Given**: TF motifs and DNase-Seq reads
- **Do**: Predict binding sites of each TF

- Identify candidate binding sites with PWMs
- Fit Gaussian process parameters for background
- Estimate TF binding effects $\beta_{i-j,l}$
- Iterate until parameters converge
  - Estimate Gaussian process posterior with expectation propagation
  - Estimate expectation of which candidate binding sites are bound
  - Update monotonic regression functions for binding priors

TF binding hierarchy

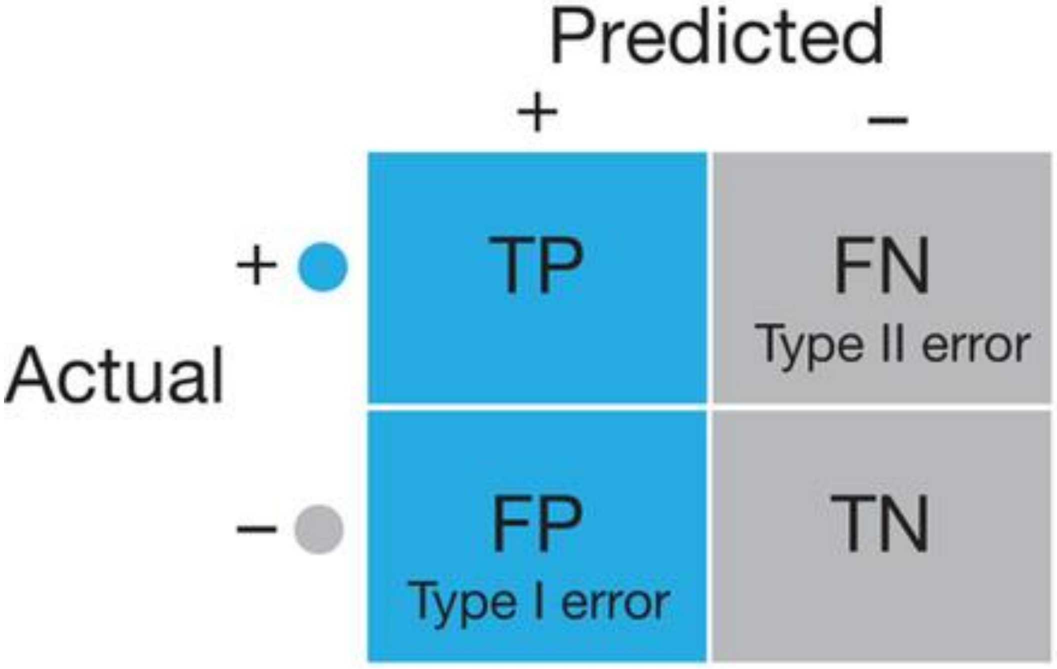
- Pioneer, settler, and migrant TFs



Sherwood *Nature Biotechnology* 2014

Evaluation: confusion matrix

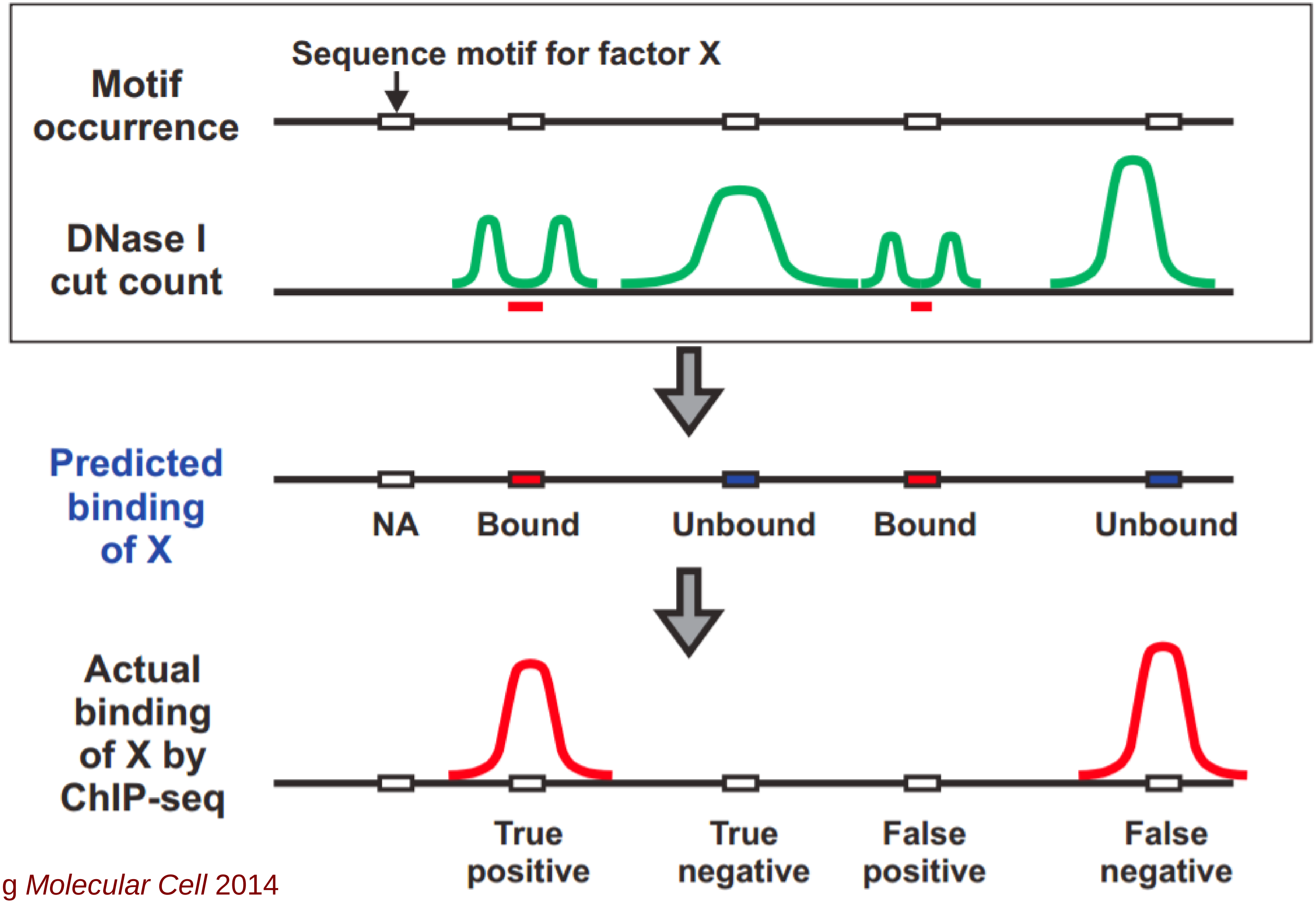
- Compare predictions to actual ground truth (gold standard)



		Predicted	
		+	–
Actual	+ ●	TP	FN Type II error
	– ●	FP Type I error	TN

Lever *Nature Methods* 2016

Evaluation: ChIP-Seq gold standard



Sung *Molecular Cell* 2014

Evaluation: ROC curve

- Calculate **r**eceiver **o**perating **c**haracteristic curve (ROC)
- True Positive Rate versus False Positive Rate
- Summarize with **a**rea **u**nder ROC **c**urve (AUC ROC)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

Includes true negatives
Reason to prefer precision-recall for class imbalanced data

Evaluation: ROC curve

- TPR and FPR are defined for a **set** of positive predictions
- Need to threshold continuous predictions
- Rank predictions
- ROC curve assesses all thresholds

Candidate binding site	*P*(bound)
764	0.99
47	0.96
942	0.91
157	0.87
79	0.83
202	0.72
356	0.66
679	0.51
291	0.43
810	0.40
…	

t

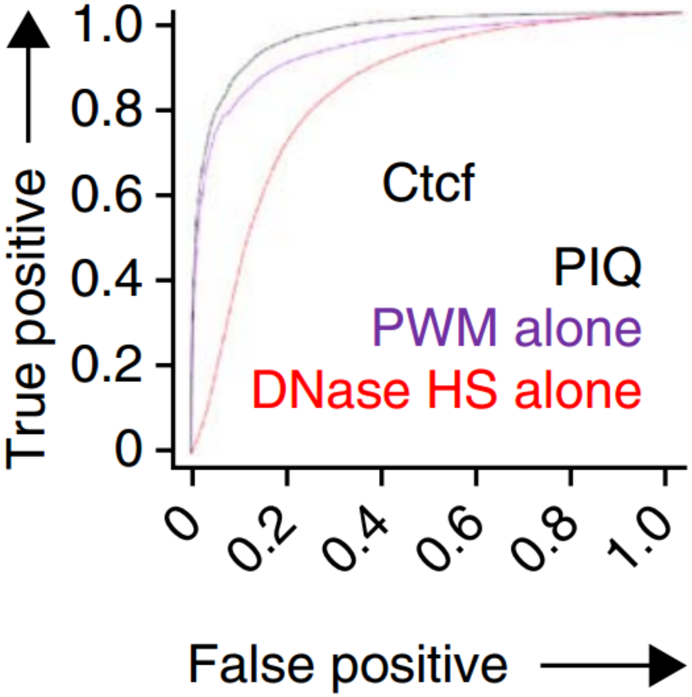
Positive predictions

Negative predictions

Calculate TPR and FPR at all thresholds *t*

PIQ ROC curve for mouse Ctcf

- Compare predictions to ChIP-Seq
- Full PIQ model improves upon motifs or DNase alone

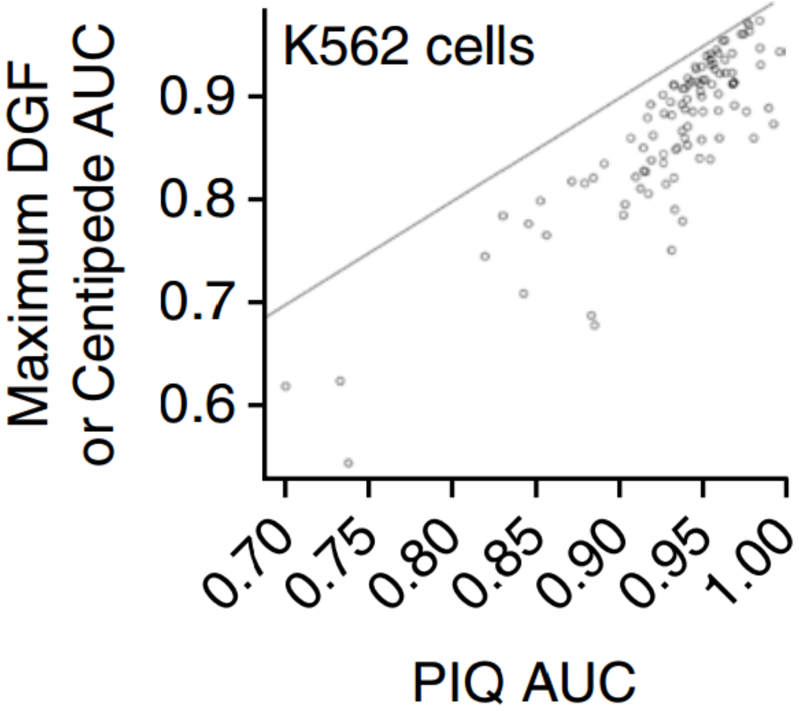


Sherwood *Nature Biotechnology* 2014

PIQ evaluation

- Compare to two standard methods
  - 303 ChIP-Seq experiments in K562 cells
  - Centipede, digital genomic footprinting

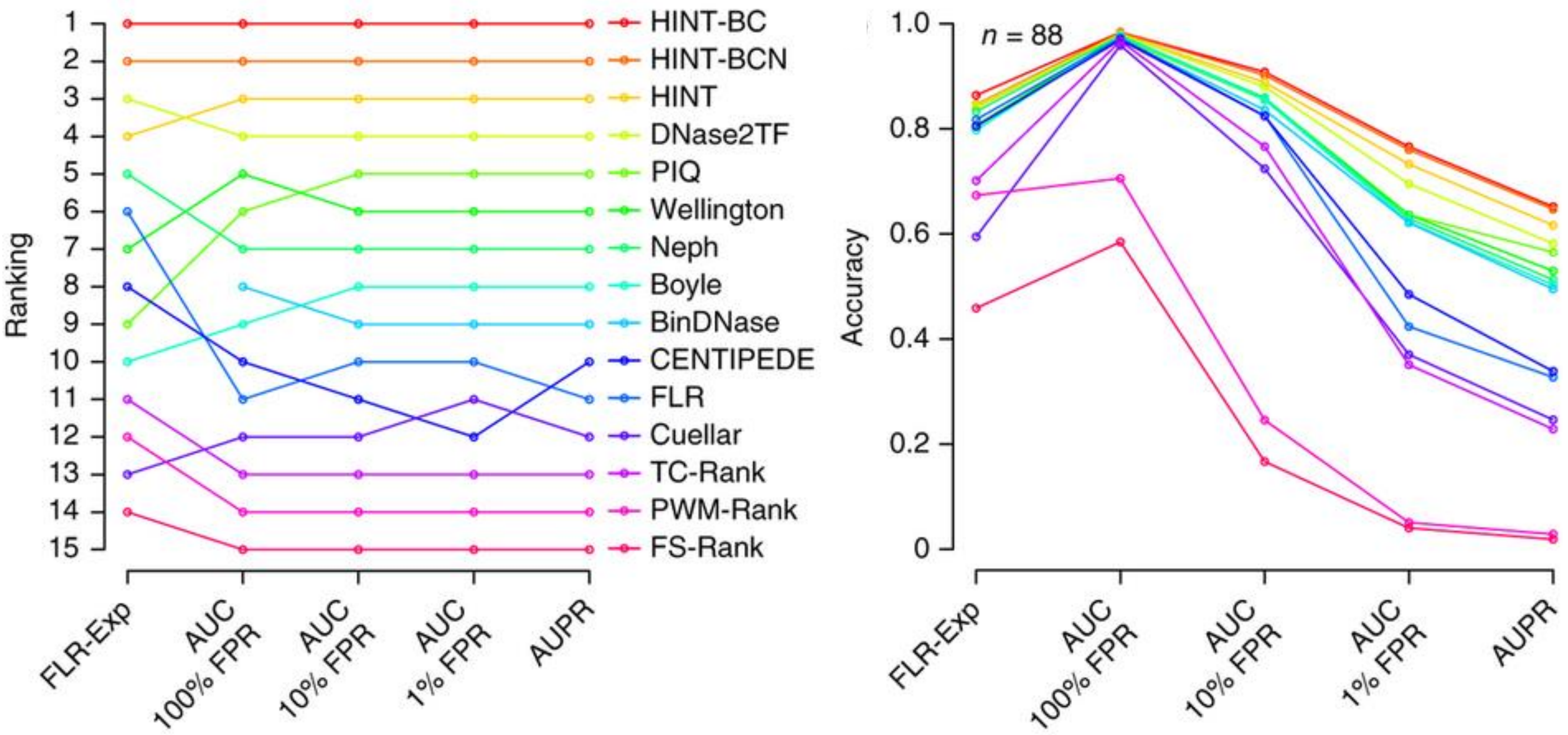
- Compare AUC ROC
  - PIQ has very high AUC
  - Mean 0.93
  - Corresponds to recovering median of 50% of binding sites



Sherwood *Nature Biotechnology* 2014

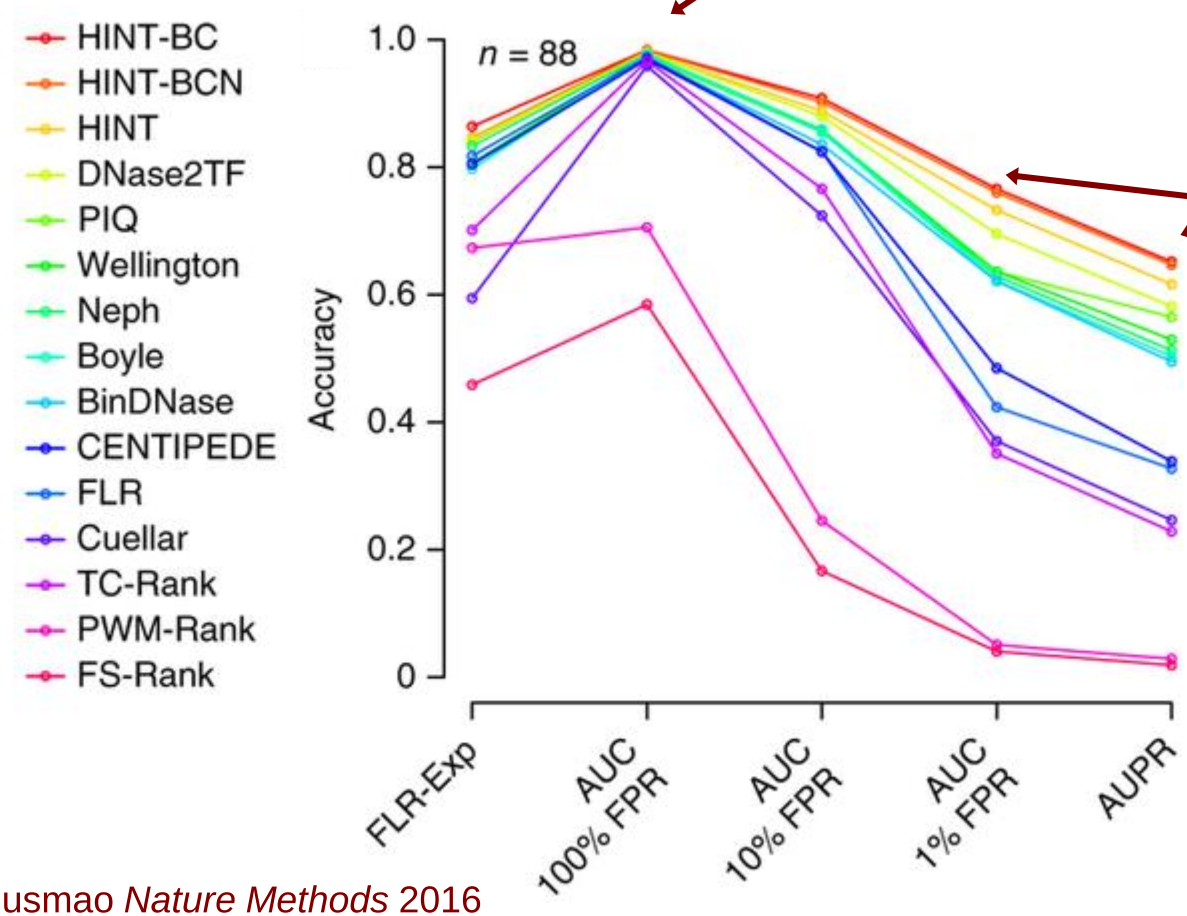
DNase-Seq benchmarking

- PIQ among top methods in large scale DNase benchmarking study
- HMM-based model HINT was top performer



Gusmao *Nature Methods* 2016

Downside of AUC ROC for genome-wide evaluations



Almost all methods look equally good when using full ROC curve AUC ROC close to 1.0

Precision-recall curve or truncated ROC curve differentiate methods

Gusmao *Nature Methods* 2016

PIQ summary

- Smooth noisy DNase-Seq data without imposing too much structure

- Combine DNase-Seq and motifs to predict condition-specific binding sites

- Supports replicates and multiple related conditions (e.g. time series)